Project 1 Proposal

1. Project Details

# Project Topic：**Urban Crime Rate Prediction**

# Team Members: Hongyu Yan, Chenshu Xu

1. Project Data Resources

### Crimes Data by Chicago Police Department - 2001 to present

<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>

Used as main database. Include all crime data from 2001 to present.

### Crimes in Chicago

<https://www.kaggle.com/currie32/crimes-in-chicago>

This is an extensive dataset of crimes in Chicago (2001-2017), can be a supplement of the main database.

### U.S. Hourly Precipitation Data

<https://catalog.data.gov/dataset/u-s-hourly-precipitation-data>

This database can be used for get auxiliary features. We plan to combine the daily weather information with main database. Because we think the weather condition can affect the crime rate.

### Average Daily Traffic Counts

<https://data.cityofchicago.org/Transportation/Average-Daily-Traffic-Counts/pfsx-4n4m>

This database is used for get auxiliary features. This database need preproces because it only have the total count of traffic passing in column, which need to be formed into the rate of traffic condition like high, medium and low.

### Census Data - Selected socioeconomic indicators in Chicago, 2008 – 2012

<https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2>

This database is used for get auxiliary features. This database need preproces because it only include the community area name in column. So we need to match each Community Area in the main database with community area names then we can merge the two databases. The column of “Percent aged 16+ unemployed”, “Percent aged 25+ without high school diploma” and “Per capita income” can be used to group the similar community and be used as a variable to predict crime rate.

1. Abstract

The goal of this project is to develop a effective predictive modeling framework with the historical crimes data, weather data, traffic data and demographic data. The predictive modeling will be used to predict the possibility that a crime will occur with a specific location in Chicago. The three variables are set in this predictive modeling which are location, time of day and day of week. The location can be a latitude and longitude, a county name or a zip code. The time of day can be one of the eight time period (00:00 - 03:00, 03:00 - 06:00, 06:00 - 09:00, 09:00 - 12:00, 12:00 - 15:00, 15:00 - 18:00, 18:00 - 21:00, 21:00 - 24:00) . The day of week is one day of the week like Monday. With these variables, the predictive modeling will predict the probability that a crime will occur.

1. Project Timeline

### Phase 1 (2/12 - 3/18)

#### Week 1 (2/12 - 2/18)

Collect the crime data of Chicago.

#### Week 2 (2/19 - 2/25)

Collect the auxiliary data of Chicago.

#### Week 3-5 (2/26 - 3/18)

Preprocessing all the datas of Chicago.

1. Check abnormal datas, make subregions of Chicago (Week 3)
2. Make sub dataframe of all datas with subregions (Week 4)
3. Combine sub dataframes with same subregions (Week 4 and 5)
4. Combine into a big Database (Week 5)

### Phase 2 (3/19 - 4/29)

1. Role of Team Members

#### Chenshu Xu:

1. Collect the the crime data of Chicago.
2. Check and remove the abnormal data in the crime data and auxiliary data.
3. Separate the crime data into sub-dataframes with sub-regions.
4. Combine those sub dataframes of the crime data and the auxiliary data into a Dataframe for each subregions.
5. Combine those DataFrames into a Database for the predictive modeling.

#### Hongyu Yan:

1. Collect the auxiliary data like weather data (weather condition), traffic data(traffic condition), demographic data(population density and income level) of Chicago.
2. Separate the Chicago into smaller sub-regions.
3. Separate the auxiliary data into sub dataframes with subregions.
4. Combine those sub dataframes of the crime data and the auxiliary data into a Dataframe for each subregions.